Genewindow: an interactive tool for visualization of genomic variation

To the editor:

To facilitate the visualization of genetic variation in the context of the human genome, we developed a tool that permits 'gene-centric' annotation of the human genome for laboratory and analytical work carried out at the Core Genotyping Facility (CGF) of the National Cancer Institute. This tool integrates data available in the public databases with internal annotations from sequence data generated by our laboratory. The Genewindow tool is now publicly available for viewing human variation and CGF sequence-validated SNPs. Although Genewindow is configured for the human genome and integrated with our laboratory data, it can be applied to other genomes and integrated with the analysis, storage and archiving of data generated in any laboratory setting.

The ability to scan and specifically examine sequence data for neighboring SNPs is critical, because common variations occur on average at least once every several hundred bases[1]. But development of tools for the visualization and manipulation of such large amounts of genetic data is daunting. It can also be challenging for investigators to organize dense annotated data sets, particularly with respect to culling disparate sources of data for the design and execution of genetic studies[2]. Previously, solutions have required labor-intensive manual annotation of genes, which can be error-prone, particularly when using multiple intermediate files and third-party software packages.

The Genewindow website was designed to represent genomic variation intuitively for analyses using advanced web-based graphics. Users can search by HUGO gene symbol, dbSNP ID, an internal CGF polymorphism ID or chromosome coordinates to view a gene or variation of interest. Genewindow is considered gene-centric only when a gene of interest is in view and the display is oriented 5′ to 3′, regardless of the reference strand and adjacent genes. The display is split into two views: at the top, a Locus Overview, which varies in size depending on the gene or genomic region being viewed, and below it, a Sequence View showing 2,000 base pairs within the overview (**Fig. 1**). The Locus Overview shows alternate translations (purple segments), current protein (purple connecting segments), regions under investigation by the CGF (green), current gene (gold), neighboring genes (brown) and polymorphisms (which can be color-coded). The vertical and diagonal lines connecting the Locus Overview and the Sequence View indicate their spatial relationship. Changing the Sequence View to show different sequence regions is as simple as clicking along the gene in the Locus Overview. The gray bar at the far left is a menu that allows users to expand or contract the genomic interval, access lists of features, search for sequence matches or view the legend. Other options for navigating a region or gene of interest include shifting the view in the 5′ or 3′ direction (relative to the current gene) or using arrow buttons on either side of the Locus Overview. Genomic features are represented by shape, color and opacity with contextual information available when the user moves over or clicks on a feature. For example, polymorphisms with low opacity indicate low frequency as defined by either dbSNP or a CGF sequence-validated assay. More explanations are give in the Genewindow legend on the website.

Administrators can insert newly discovered polymorphisms into the Genewindow database by entering annotations directly through the graphical user interface or through an automated pipeline. Genewindow generates export files that include the target polymorphism and its surrounding annotations, which are imported into both our Laboratory Information Management System and our public website of sequence validated

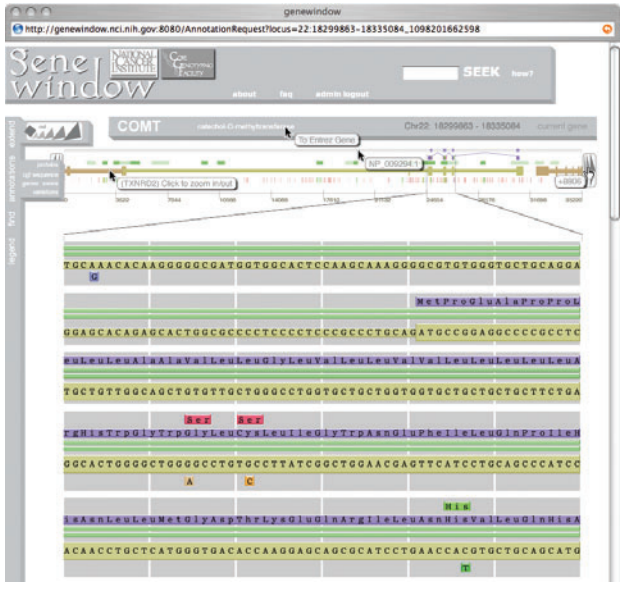

Figure 1 Genewindow display: Locus Overview on top, Sequence View below.

SNPs, SNP500Cancer[3]. The Genewindow website is now the primary tool for pre- and postgenetic bioinformatics and analytical work at the CGF, which currently has more than 75,000 study samples available, and last year delivered more than four million genotypes.

The Genewindow tool is a work in progress. Future plans include the addition of features essential for the study of candidate genes and human variation, namely the abilities to toggle the display of polymorphisms based on data sources and to download features and data in a desired region. For example, it will be possible to highlight SNPs identified by the International HapMap[4]. Additional features will include the display and export of estimated haplotypes, haplotype-tagging SNPs and linkage disequilibrium. Further public data imports will include regulatory and repeat regions, whole-genome human-mouse alignments and gene ontological information. These features can assist in the selection of variants for study *in vitro* or in new genetic association studies. Additionally, for the purpose of enhancing workflow and analysis in other laboratories, the application, source code and documentation will be made available for download in the near future, and thus could be tailored to meet specific laboratory needs. This can assist laboratories in choosing and tracking information related to genetic annotations. The front web page has links to supplemental information, web browser plug-in requirements and upcoming features.

URLs. The Genewindow tool is available at http://genewindow.nci.nih.gov/. Our public website of sequence validated SNPs, SNP500Cancer, is available at http://snp500cancer.nci.nih.gov/.

Brian Staats[1,2], Liqun Qi[1,2], Michael Beerman[1,2], Hugues Sicotte[1,2], Laura A Burdett[1,2], Bernice Packer[1,2], Stephen J Chanock[1,3] & Meredith Yeager[1,2]

[1]Core Genotyping Facility, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Gaithersburg, Maryland 20877, USA. [2]Intramural Research Support Program, SAIC-Frederick, NCI-FCRDC, Frederick, Maryland 21702, USA. [3]Section on Genomic Variation, Pediatric Oncology Branch, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. Correspondence should be addressed to M.Y. (yeagerm@mail.nih.gov).

1. Carlson, C.S. *et al. Am. J. Hum. Genet.* **74**, 106–120 (2004).
2. Risch, N.J. *Nature* **405**, 847–856 (2000).
3. Packer, B.R. *et al. Nucleic Acids Res.* **32**, D528–D532 (2004).
4. International HapMap Consortium. *Nature* **426**, 789–796 (2003).

Assessing the validity of the association between the SUMO4 M55V variant and risk of type 1 diabetes

To the editor:

Two groups recently reported evidence for association of a common nonsynonymous SNP (163A→G or rs237025, resulting in the amino acid substitution M55V) in the gene small ubiquitin-related modifier 4 (*SUMO4*) with type 1 diabetes (T1D)[1–3]. But an inconsistency is evident in a comparison of their results: Bohren *et al.*[1] claimed that the 163A allele was positively associated with disease susceptibility, whereas Guo *et al.*[2] reported that the nonconserved 163G allele was causal. We noted that in Guo *et al.*'s study of 944 families, the transmissions for these alleles in 92 UK multiplex families (from the Diabetes UK Warren repository) were in the opposite direction to the rest of their family results and case-control analyses, with the 163A allele transmitted at 56% (149 transmissions), whereas the 163G allele was transmitted at 58.6% in the rest of the families (922 transmissions). The inclusion of the 92 UK multiplex families had a large effect on the *P* value reported by Guo *et al.* ($P = 1.9 \times 10^{-7}$ to 2.9×10^{-5}). Moreover, their results from these 92 UK families were consistent with the increased transmission of the 163G allele observed by Bohren *et al.*, who analyzed 222 UK families from the Warren repository (which probably included the 92 families studied by Guo *et al.*) and 256 European American families (from the Human Biological Data Interchange collection)[1,3].

Table 1 Transmission disequilibrium test results for the *SUMO4* 163A→G SNP

Population	Number of parent-child trios	Parental allele frequency	Transmitted (A)	Untransmitted (G)	%T[a]	P
UK[b]	825	0.49/0.51	418	376	53	0.14
USA	609	0.49/0.51	281	306	48	0.30
N. Ireland	248	0.49/0.51	120	129	48	0.56
Finland	829	0.54/0.46	431	412	51	0.51
Romania	223	0.50/0.50	120	101	54	0.20
Norway	317	0.53/0.47	166	154	51	0.57
All	3,051	0.51/0.49	1,534	1,473	51	0.31

[a]Percentage transmission. [b]Includes families from Yorkshire.

We, therefore, genotyped the 163A→G SNP in similar, but expanded, collections of families from the UK Warren repository ($n = 471$) and the US Human Biological Data Interchange ($n = 336$; **Table 1** and **Supplementary Methods** online). We did not obtain any evidence of association with T1D in 1,381 transmissions (50.6 % for the 163A allele). Guo *et al.* obtained additional support for the 163G association in 314 cases and controls from Finland, but in a much larger sample of 911 families from Finland, we found no evidence of association (51% transmission of the 163A allele; **Table 1**). In additional families from Northern Ireland, Yorkshire, Romania and Norway, there was also no evidence of association (**Table 1**). Finally, we analyzed a large case-control sample from Great Britain (**Supplementary Methods** online) consisting of 3,442 cases (from the UK GRID) and 3,788 population-based controls (from the 1958 British Birth Cohort) and obtained virtually identical allele frequencies in the cases and controls (0.49 for the 163A allele; **Supplementary Table 1** online). To confirm the validity of our assay, we resequenced *SUMO4* in 32 individuals and observed no discrepancies (we did identify a new, rare SNP (ss28525188) that changes the SUMO4 amino acid sequence;